Exploratory Data Analysis using Logistic Regression

**By:**

**Team Tech Tycoons**

B.krishna Chaitanya

Vyshnavi

Sadhana

Kavitha

Bhanu Prakash

Tulasi

**Contents**

Introduction

* Machine Learning………………………………..3
* Python……………………………………………6

Objectives of work

Problem Statement………………………………………14

Review of Literature…………………………………….15

Data Collection………………………………………….14

Methodology

* **Exploratory data analysis**

1. Distplot for

environmental satisfaction………………..19

Job Involvement………………………….19

Years with curr manager…………………20

Years in current role……………………..20

* **2.Box plots for**

Environmental Satisfaction………………..21

Total Working Years………………………21

Over Time…………………………………21

Years in current role………………………22

Job Involvement………………………….22

Stock Option Level………………………22

Data Normalization……………………………………23

Logistic Regression……………………………………24

Models…………………………………………………27

Findings and Suggestions……………………………..30

Conclusion…………………………………………….31

References…………………………………………….31

**Introduction**

**Machine learning**

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

**Machine learning tasks**

Machine learning tasks are classified into several broad categories as supervised and unsupervised.

**Supervised machine learning algorithms**

In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs.

 For example, if the task were determining whether an image contained a certain object, the training data for a supervised learning algorithm would include images with and without that object (the input), and each image would have a label (the output) designating whether it contained the object. In special cases, the input may be only partially available, or restricted to special feedback.

**Unsupervised learning machine learning algorithms**

is a term used for Hebbian learning, associated with learning without a teacher, also known as self-organization and a method of modelling the probability density of inputs?

Unsupervised machine learning algorithms infer patterns from a dataset without reference to known, or labelled, outcomes. Unlike supervised machine learning, unsupervised machine learning methods cannot be directly applied to a regression or a classification problem because you have no idea what the values for the output data might be, making it impossible for you to train the algorithm the way you normally would. Unsupervised learning can instead be used for discovering the underlying structure of the data.

**Reinforcement machine learning algorithms**

Reinforcement learning, in the context of artificial intelligence, is a type of dynamic programming that trains algorithms using a system of reward and punishment.

A reinforcement learning algorithm, or agent, learns by interacting with its environment. The agent receives rewards by performing correctly and penalties for performing incorrectly. The agent learns without intervention from a human by maximizing its reward and minimizing its penalty.

**Types of Learning**

There are four types of machine learning:

**Supervised learning:** (also called inductive learning) Training data includes desired outputs.  This is spam this is not, learning is supervised.

**Unsupervised learning:** Training data does not include desired outputs. Example is clustering. It is hard to tell what is good learning and what is not.

**Semi-supervised learning:** Training data includes a few desired outputs.

**Reinforcement learning:** Rewards from a sequence of actions. AI types like it, it is the most ambitious type of learning.

**Terminology**

To avoid confusion due to ambiguity, here are some definitions of terms used in this book

**Machine Learning** is a set of methods that allow computers to learn from data to make and improve predictions (for example cancer, weekly sales, credit default). Machine learning is a paradigm shift from “normal programming” where all instructions must be explicitly given to the computer to “indirect programming” that takes place through providing data.

**Models**

A model defines the relationship between features and label. For example, a spam detection model might associate certain features strongly with "spam". Let's highlight two phases of a model's life:

**Target(Label)**A target variable or label is the value to be predicted by our model. For the fruit example discussed in the features section, the label with each set of input would be the name of the fruit like apple, orange, banana, etc.

**Training**

means creating or **learning** the model. That is, you show the model labelled examples and enable the model to gradually learn the relationships between features and label.

**Inference**

means applying the trained model to unlabelled examples. That is, you use the trained model to make useful predictions (y'). For example, during inference, you can predict medianHouseValue for new unlabelled examples.

The figure shown below clears the above concepts:

**Modeling**

Exploratory Data Analysis in Python

What is Exploratory Data Analysis (EDA) ?

EDA is a phenomenon under data analysis used for gaining a better understanding of data aspects like:  
–main features of data  
– variables and relationships that hold between them  
– identifying which variables are important for our problem  
We shall look at various exploratory data analysis methods like:

* Descriptive Statistics, which is a way of giving a brief overview of the dataset we are dealing with, including some measures and features of the sample.
* Grouping data [Basic grouping with group by]
* ANOVA, Analysis Of Variance, which is a computational method to divide variations in an observations set into different components.
* Correlation and correlation methods

**Python Introduction**

**Python** is a general purpose, dynamic, high level and interpreted programming language. It supports Object Oriented programming approach to develop applications. It is simple and easy to learn and provides lots of high-level data structures.

Python has become the lingua franca for many data science applications. It combines the power of general-purpose programming languages with the ease of use of domain-specific scripting languages like MATLAB or R. Python has libraries for data loading, visualization, statistics, natural language processing, image processing, and more.

This vast toolbox provides data scientists with a large array of general- and special-purpose functionality.

One of the main advantages of using Python is the ability to interact directly with the code, using a terminal or other tools like the Jupyter Notebook, which we’ll look at shortly.

Machine learning and data analysis are fundamentally iterative processes, in which the data drives the analysis.

It is essential for these processes to have tools that allow quick iteration and easy interaction.

As a general-purpose programming language, Python also allows for the creation of complex graphical user interfaces (GUIs) and web services, and for integration into existing systems.

**Python History**

* Python laid its foundation in the late 1980s.
* The implementation of Python was started in the December 1989 by Guido Van Rossum at CWI in Netherland.
* In February 1991, van Rossum published the code (labeled version 0.9.0) to alt.sources.
* In 1994, Python 1.0 was released with new features like: lambda, map, filter, and reduce.
* Python 2.0 added new features like: list comprehensions, garbage collection system.
* On December 3, 2008, Python 3.0 (also called "Py3K") was released. It was designed to rectify fundamental flaw of the language.
* *ABC programming language* is said to be the predecessor of Python language which was capable of Exception Handling and interfacing with Amoeba Operating System.
* Python is influenced by following programming languages:

ABC language.

Modula-3

**Python Features**

Python provides lots of features that are listed below.

**1) Easy to Learn and Use**

Python is easy to learn and use. It is developer-friendly and high level programming language.

**2) Expressive Language**

Python language is more expressive means that it is more understandable and readable.

**3) Interpreted Language**

Python is an interpreted language i.e. interpreter executes the code line by line at a time. This makes debugging easy and thus suitable for beginners.

**4) Cross-platform Language**

Python can run equally on different platforms such as Windows, Linux, Unix and Macintosh etc. So, we can say that Python is a portable language.

**5) Free and Open Source**

Python language is freely available at offical web address.The source-code is also available. Therefore it is open source.

**6) Object-Oriented Language**

Python supports object oriented language and concepts of classes and objects come into existence.

**7) Extensible**

It implies that other languages such as C/C++ can be used to compile the code and thus it can be used further in our python code.

**8) Large Standard Library**

Python has a large and broad library and prvides rich set of module and functions for rapid application development. And widely supports Machine Learning because of rich libraries and packages.

**Python Variables**

* Variable is a name which is used to refer memory location.
* Variable also known as identifier and used to hold value.
* In Python, we don't need to specify the type of variable because Python is a type infer language and smart enough to get variable type.
* Variable names can be a group of both letters and digits, but they have to begin with a letter or an underscore.
* It is recomended to use lowercase letters for variable name. Rahul and rahul both are two different variables.

Python has five standard data types –

A variable can hold different types of values. For example, a person's name must be stored as a string whereas its id must be stored as an integer.

Python provides various standard data types that define the storage method on each of them. The data types defined in Python are given below.

1. Numbers
2. String
3. List
4. Tuple
5. Dictionary

**Numbers**

* Number stores numeric values. Python creates Number objects when a number is assigned to a variable.

Python supports 4 types of numeric data.

int (signed integers like 10, 2, 29, etc.)

long (long integers used for a higher range of values like 908090800L, -0x1929292L, etc.)

float (float is used to store floating point numbers like 1.9, 9.902, 15.2, etc.)

complex (complex numbers like 2.14j, 2.0 + 2.3j, etc.)

**String**

* The string can be defined as the sequence of characters represented in the quotation marks. In python, we can use single, double, or triple quotes to define a string.
* String handling in python is a straightforward task since there are various inbuilt functions and operators provided.
* In the case of string handling, the operator + is used to concatenate two strings as the operation "hello"+" python"returns "hello python".
* The operator \* is known as repetition operator as the operation "Python " \*2 returns "Python ".

**List**

* Lists are similar to arrays in C.
* However; the list can contain data of different types.
* The items stored in the list are separated with a comma (,) and enclosed within square brackets [].
* We can use slice [:] operators to access the data of the list.
* The concatenation operator (+) and repetition operator (\*) works with the list in the same way as they were working with the strings.

**Tuple**

* A tuple is similar to the list in many ways.
* Like lists, tuples also contain the collection of the items of different data types.
* The items of the tuple are separated with a comma (,) and enclosed in parentheses ().
* A tuple is a read-only data structure as we can't modify the size and value of the items of a tuple.

**Dictionary**

* Dictionary is an ordered set of a key-value pair of items.
* It is like an associative array or a hash table where each key stores a specific value.
* Key can hold any primitive data type whereas value is an arbitrary Python object.
* The items in the dictionary are separated with the comma and enclosed in the curly braces {}.

**Python Functions**

* Functions are the most important aspect of an application.
* A function can be defined as the organized block of reusable code which can be called whenever required.
* Python allows us to divide a large program into the basic building blocks known as function.
* The function contains the set of programming statements enclosed by {}.
* A function can be called multiple times to provide reusability and modularity to the python program.
* In other words, we can say that the collection of functions creates a program. The function is also known as procedure or subroutine in other programming languages.
* Python provide us various inbuilt functions like range() or print(). Although, the user can create its functions which can be called user-defined functions.

**Advantage of functions in python**

There are the following advantages of C functions.

* By using functions, we can avoid rewriting same logic/code again and again in a program.
* We can call python functions any number of times in a program and from any place in a program.
* We can track a large python program easily when it is divided into multiple functions.
* Reusability is the main achievement of python functions.
* However, Function calling is always overhead in a python program.

**Function calling**

In python, a function must be defined before the function calling otherwise the python interpreter gives an error. Once the function is defined, we can call it from another function or the python prompt. To call the function, use the function name followed by the parentheses.

**Python Classes**

Python has been an object-oriented language since it existed. Because of this, creating and using classes and objects are downright easy.

**Class:** A user-defined prototype for an object that defines a set of attributes that characterize any object of the class. The attributes are data members (class variables and instance variables) and methods, accessed via dot notation.

**Class variable:** A variable that is shared by all instances of a class. Class variables are defined within a class but outside any of the class's methods. Class variables are not used as frequently as instance variables are.

**Data member:** A class variable or instance variable that holds data associated with a class and its objects.

**Instance variable:** A variable that is defined inside a method and belongs only to the current instance of a class.

**Objectives of work**

The objective of the project is to find out if there any relation in between given columns and attrition.

If there exists any such relation how they must take actions in order to make less employee attrition.

**Problem Statement**

The given data set is about employee attrition . Where the employee attrition means that employee leaving the company by various reasons , the reasons can be personal or in the interest of company the employee might have been dismissed out of his duties . The main goal of this project is to use logistic regression to gain the knowledge out of the data .

As the process goes on we prepare models of various kind in order explain the relation ship in between the inputs and outputs .

**Data Collection :**

Data collection is the systematic approach to gathering and measuring information from a variety of sources to get a complete and accurate picture of an area of interest. Data collection enables a person or organization to answer relevant questions, evaluate outcomes and make predictions about future probabilities and trends.

The provided data is corporate data set regarding employees and they explain each and every property of each employee which can effect the employee attrition.

The data was collected from some unkown sources and is related to some anonymous corporate company.

**Review of literature**

The flow of this project work is as follows

* Importing the data into python.
* Doing Exploratory data analysis
* By using various methods of visualizations
* If the given data consists of any missing values they must be filled with some strategies.
* The above step is known as missing value treatment.
* Mostly any data is non normal , there by we must apply some normalization techniques and then we can use then data for further processes.
* The data normalization reduces varience in the data.
* Then apply the Logistic Regression on the data by splitting the data.
* There by we prepare a various kinds of models in the whole modelling . So, we later on go on to choose the models with good accuracy and other metrics.

**Exploratory Data Analysis**

The EDA generally talks about viewing the data in various dimensional views and then analysis of data by viewing at and also by preparing some predefined model.

As a part of EDA in the project we have gained some insights on the employee attrition dataset.

Dataset speaks about the various factors that effect the employee attrition and their by concluding that attrition is the output column and the rest 34 mentioned are input columns.

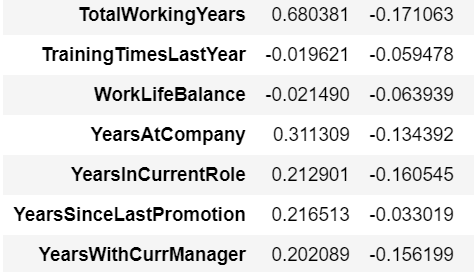
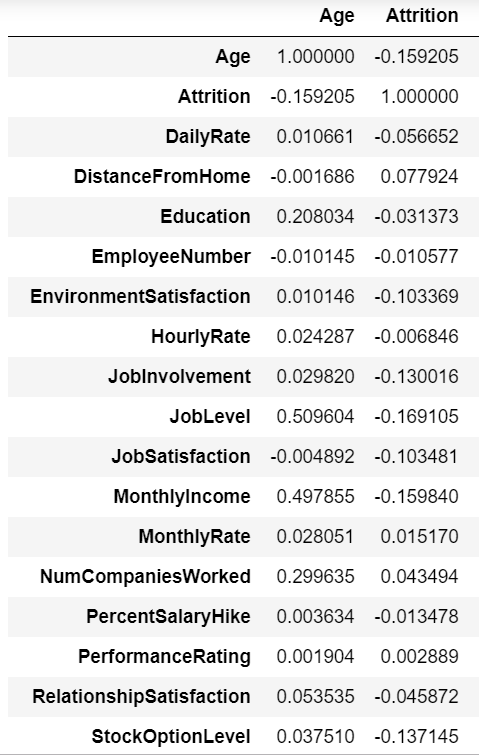
All the columns of the data set are as given below :

'Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department','DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount','EmployeeNumber', 'EnvironmentSatisfaction', 'Gender', 'HourlyRate','JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction','MaritalStatus', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked','Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating','RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel','TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance','YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion','YearsWithCurrManager'.

There are presence of categorical data columns , in order to apply Logistic Regression we have to convert the data into numerical format . And there are no missing values in the data .

And out of all those above mentioned columns we have built various kinds of visualization charts inorder to understand the given data set and then moving on we have got down to only 15 columns out of all the input columns .

The columns that we choose after considering coorelation values and by using visualization techniques like boxplots , distplots etc we have tried to understand the varience in data and the behaviours the different columns are exhibiting.



Fig(1)

**Correlation values of attrition with other inputs**

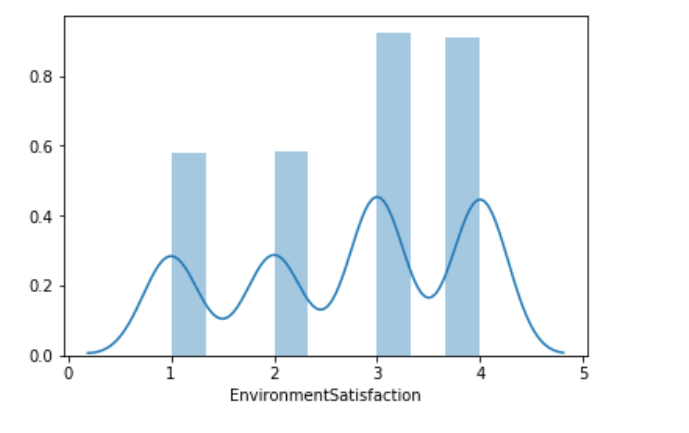


Fig.1

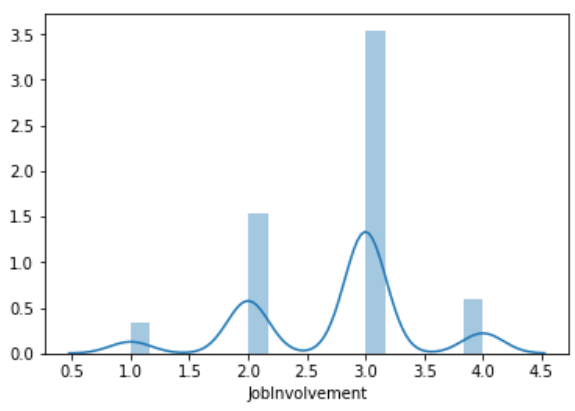
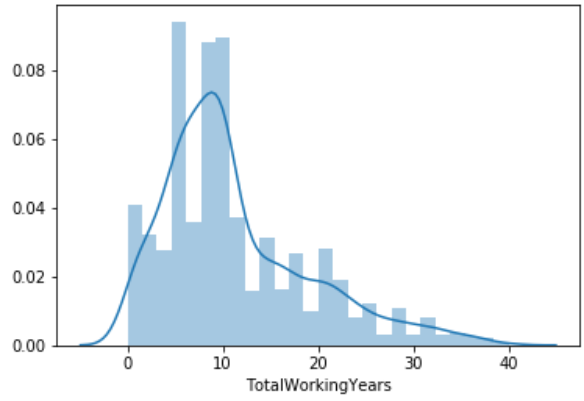


Fig.2

**Distplots for few selected colu****mn**

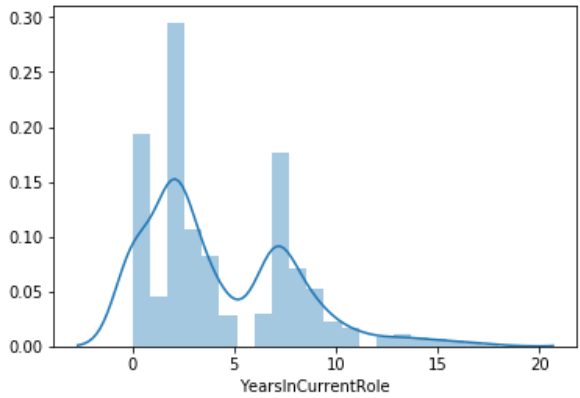


Fig.3

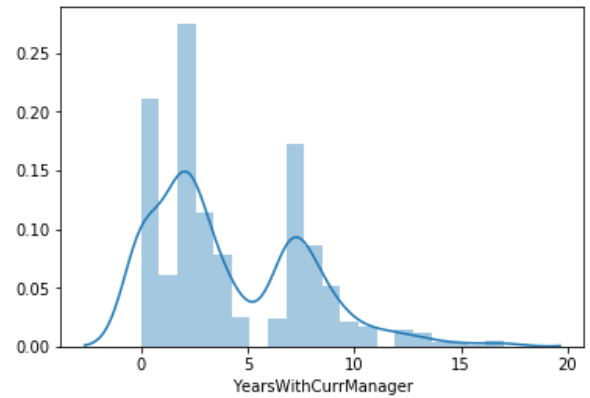


Fig.4

**Boxplots for few selected inputs**

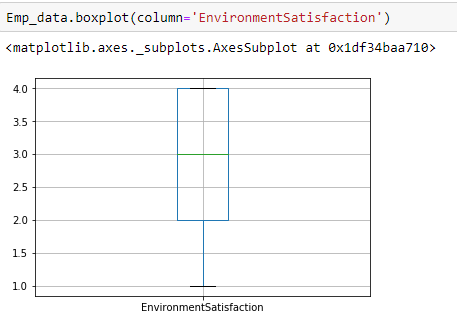


Fig.5

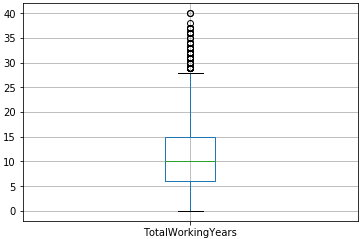


Fig.6

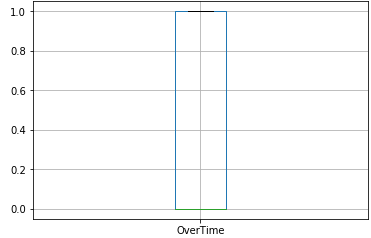


Fig.7

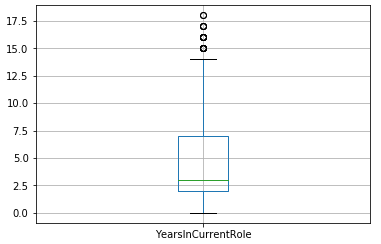


Fig.8

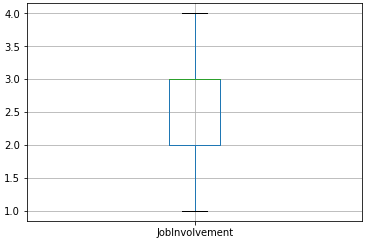


Fig.9

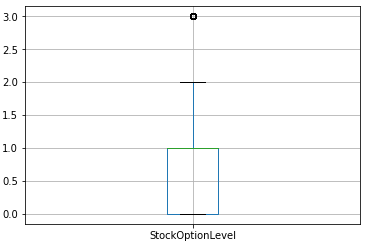
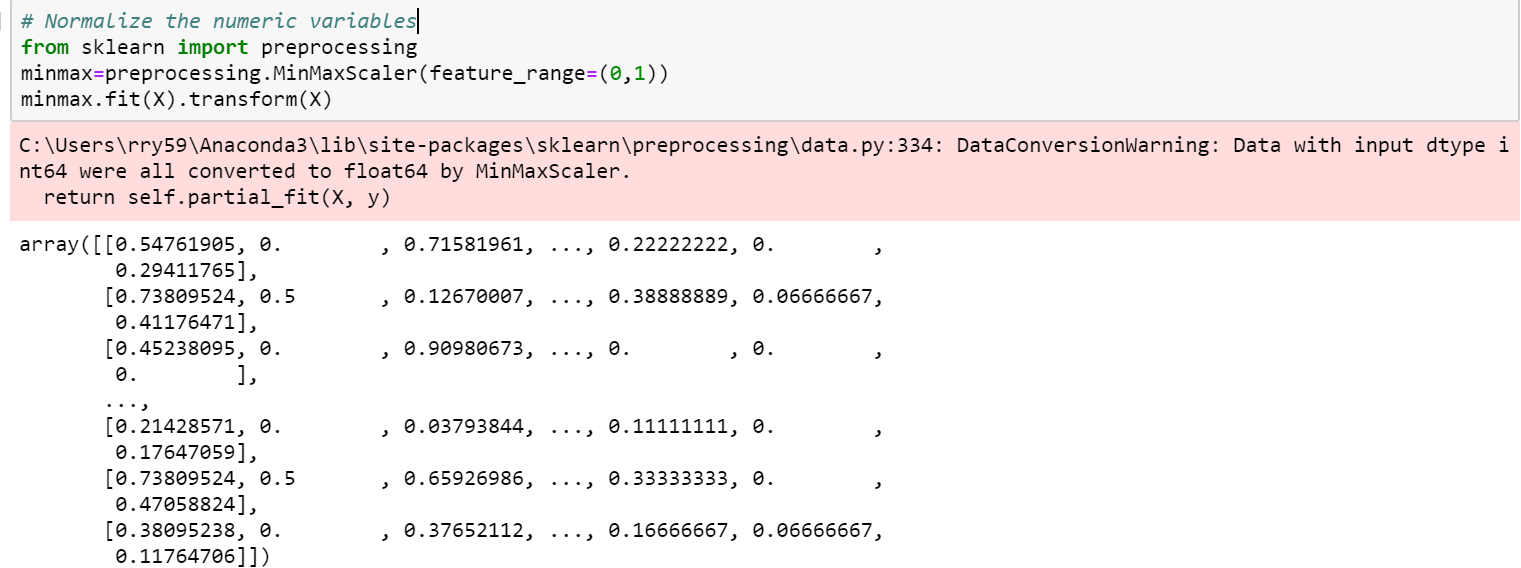


Fig.10

**Data Normalization :**

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require normalization. It is required only when features have different ranges.

For example, consider a data set containing two features, age(x1), and income(x2). Where age ranges from 0–100, while income ranges from 0–20,000 and higher. Income is about 1,000 times larger than age and ranges from 20,000–500,000. So, these two features are in very different ranges. When we do further analysis, like multivariate linear regression, for example, the attributed income will intrinsically influence the result more due to its larger value. But this doesn’t necessarily mean it is more important as a predictor.



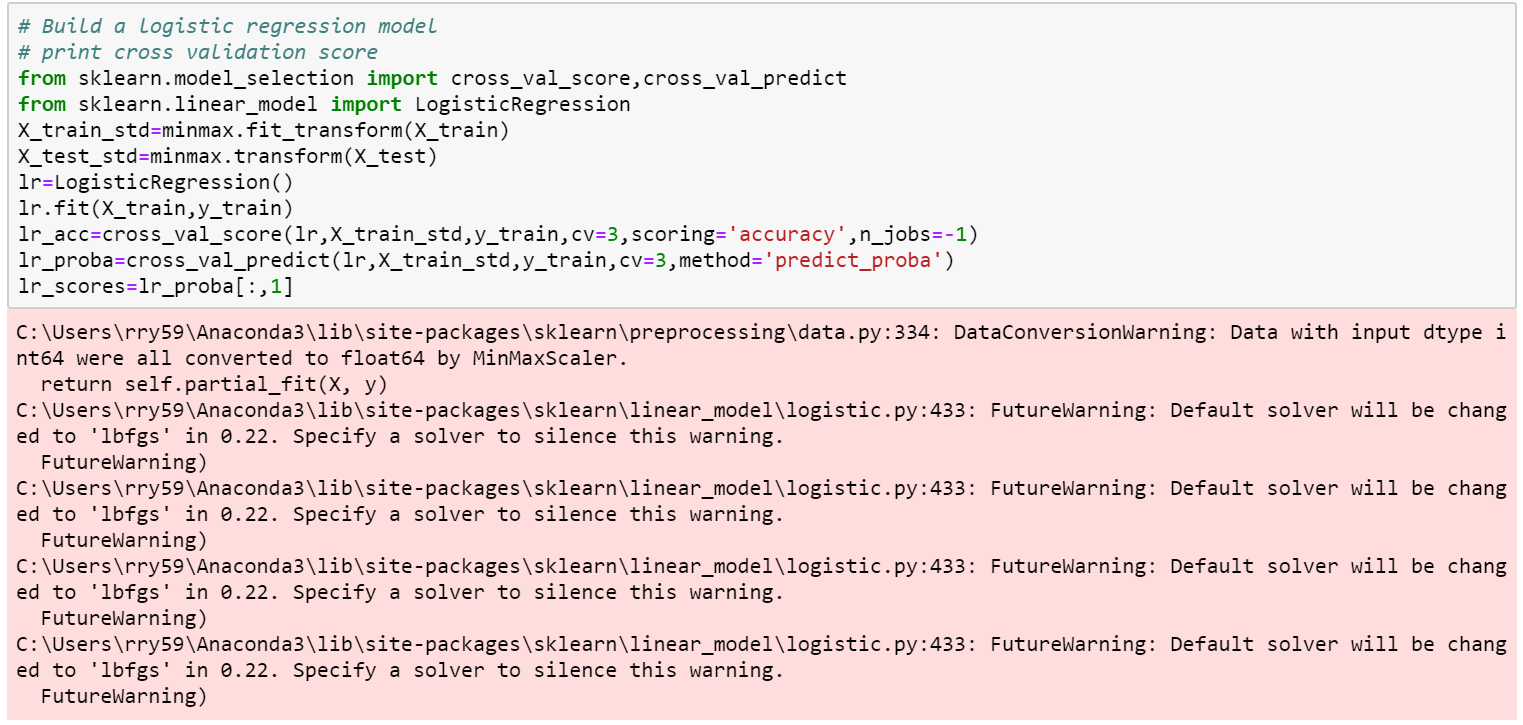
Here , the minmax normalization technique is being used for the numerical data i.e, input columns are normalized using the minmax technique of normalization.

**Logistic Regression**

Logistic regression is another technique borrowed by machine learning from the field of statistics.

It is the go-to method for binary classification problems (problems with two class values). In this post you will discover the logistic regression algorithm for machine learning.

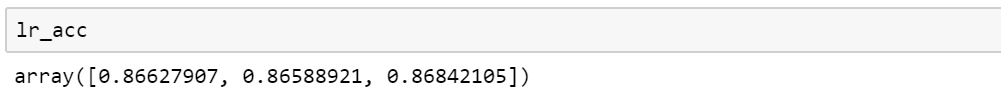
Logistic regression is the most famous machine learning algorithm after linear regression. In a lot of ways, linear regression and logistic regression are similar. But, the biggest difference lies in what they are used for. Linear regression algorithms are used to predict/forecast values but logistic regression is used for classification tasks.



The logistic regression model prepared for the model uses standard deviation values . The model is built based on difference from the mean .

We create three folds in order to explain about the accuracy of model in detail . The three folds are train ,validation and test .

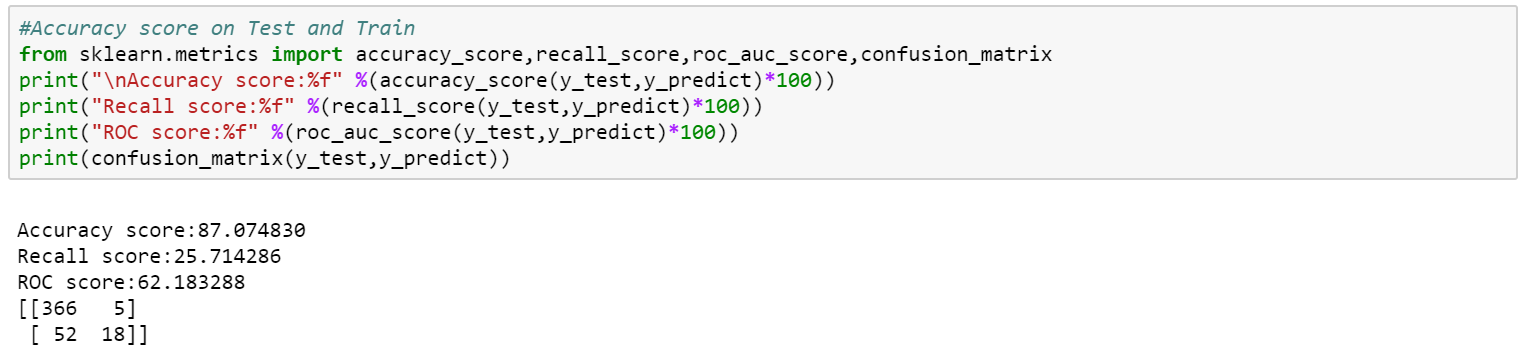
The accuracy score of three folds is calculated using cross validation .



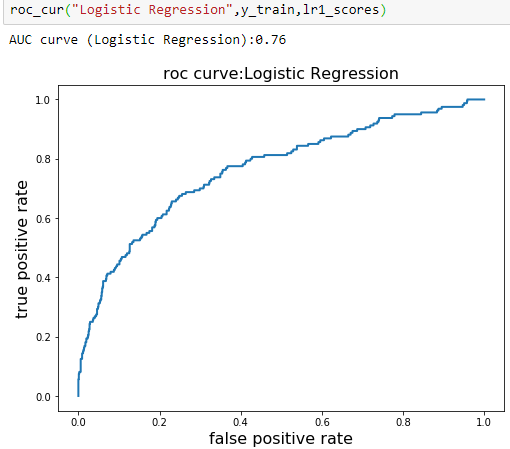
**Other metrics**

From sklearn we have imported accuracy\_score , roc\_auc\_score , confusion\_matrix .

* The accuracy score talks about accuracy of model on test data
* The Recall score speaks about the true positive rate of the model
* The roc\_auc\_score is about the sensitivity and the specificity .

****

**ROC curves**



Generally the ROC curves of the model explains about the true positive rate and the false positive rate of the model .

The AUC refers to area under the curve . If the value is close to 0 then model is said to be poor . If the value is close 0.5 model is said to be average . If the value is closer to 1 then model is peforming well .

**Models:**

**Single variant logistic regression accuracy , test size=0.3**

Years in current role = 86..84

Years with current manager = 83.67

Environment satisfaction = 84.12

Job involvement = 82.31

Job satisfaction = 85.03

Total years work = 81.85

Overtime =82.53

Stock option level = 81.52

**Multi variant logistic regression accuracy , test size=0.3**

environment satisfaction , job involvement = 82.53

total working years , years with current manager= 80.04

job satisfaction , years in current role = 82.76

over time , stock option = 82.08

stock option level , total working years , years in current role , years with current manager =83.67

job satisfaction, over time, total working years, stock option level =83.44

environmental satisfaction, job involvement, years in current role, years with curr manager=84.58

environment satisfaction , job involvement ,job satisfaction , overtime =83.4

,environment satisfaction , job involvement ,job satisfaction , overtime , stock option level , total working years , years in current role , years with current manager = 84.35

Department , distance from home , education field , environment satisfaction , job involvement ,job satisfaction , number of companies worked , overtime ,percent salary hike ,stock option level , total working years , years in current role , years with current manager = 84.35

Complete data set accuracy = 85.94

**Single variant logistic regression accuracy, test size = 0.2**

Environmental satisfaction = 84.35

Job involvement = 79.93

Job satisfaction = 84.01

Stock option level= 84.35

Over time = 82.65

Years in current role =84.35

Years with current manager = 83.33

Total work years=81.29

**Multi variant logistic regression accuracy, test size = 0.2**

job satisfaction , years in current role = 81.97

environment satisfaction , job involvement = 80.04

total working years , years with current manager= 86.39

over time , stock option = 82.08

stock option level ,total working years ,years in current role , years with current manager=79.93

job satisfaction, over time, total working years, stock option level =86.73

environmental satisfaction, job involvement, years in current role, years with curr manager=81.29

environment satisfaction , job involvement ,job satisfaction , overtime , stock option level , total working years , years in current role , years with current manager = 82.6

Department , distance from home , education field , environment satisfaction , job involvement ,job satisfaction , number of companies worked , overtime ,percent salary hike ,stock option level , total working years , years in current role , years with current manager = 86.39

Complete data set accuracy = 85.71

**The Best Performing models that we have choosen :**

**Single variant model accuracy :**

Years in current role , test size(0.3) = 86..84

Job satisfaction ,test size(0.3)= 85.03

Stock option level, test size(0.2)= 84.35

Environment Satisfaction , test size(0.2) = 84.35

**Multi variant model accuracy:**

total working years , years with current manager , test size(0.2)= 86.39

job satisfaction , years in current role = 82.76

job satisfaction, over time, total working years, stock option level =86.73

environmental satisfaction, job involvement, years in current role, years with curr manager=84.58

Department , distance from home , education field , environment satisfaction , job involvement ,job satisfaction , number of companies worked , overtime ,percent salary hike ,stock option level , total working years , years in current role , years with current manager = 84.35

**Findings and Suggestions**

* There are few unwanted or uninfluential columns present in the data
* The Years is current role has played crutial role in employee attirition according to our models .
* The data looks like it is noisy because of its outliers and the non normal from .
* In consideration to stat models and correlation data , we have selected 8 crutial inputs , they are environment satisfaction , job involvement , years in current role , stock option level , over time , years with current manager
* All these inputs have effect on attrition , there I would predict that employee attrition is mainly because of the employee personal reasons rather than company dismissing employee .
* In such case , I would suggest to improve few conditions and situations in the company based on which these inputs will be effected and result in changing attrition.

**Conclusion**

The models created are mainly to identify if there exists any relation how they are related .

Which variables effect the output must be found and mainly extract some essential knowledge out of the data.

The data on which Machine learning done must be made sutiable for such operations.

The techniques and algorithms must be adopted accordingly.

The problem must be correctly identified as classification or regression.

**References**

<https://hackernoon.com/introduction-to-machine-learning-algorithms-logistic-regression-cbdd82d81a36>

<https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148>